DOCUMENT RESUME

ED 306 254 TM 013 048

AUTHOR Cason, Gerald J.; Cason, Carolyn L.
TITLE Rater Stringency Error in Performance Rating: A Contrast of Three Models.
PUB DATE 89
NOTE 26p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; Achievement Rating; *Error of Measurement; Evaluation Methods; Higher Education; *Interrater Reliability; *Mathematical Models; Medical Students; *Performance; Rating Scales; Sample Size
IDENTIFIERS *Performance Rating Theory (Cason and Cason); Rater Response Theory; *Rater Stringency Error

ABSTRACT

The use of three remedies for errors in the measurement of ability that arise from differences in rater stringency is discussed. Models contrasted are: (1) Conventional; (2) Handicap; and (3) deterministic Rater Response Theory (RRT). General model requirements, power, bias of measures, computing cost, and complexity are contrasted. Contrasts are illustrated by application of the models to small and large sets of clinical performance ratings of junior-year medical students in an internal medicine clerkship in 1982. The small sample consisted of the ratings for a cohort of 24 students by 42 raters; 129 ratings were used. The large set included ratings for all of the students of that year; 744 ratings by 93 raters on 163 students were used. The RRT was the most powerful model, with Handicap a close second, although biased. The RRT obtained any specified interrater reliability with only a third of the independent ratings needed by the Conventional model. Pass/fail decisions are illustrated for the 163 students and a hypothetical individual student whose fate is shown to reflect his ability, the measurement model, the passing score, and the stringency and number of raters. The choice of measurement procedures should reflect balance in improving the accuracy of the measures and the costs of faulty decisions based on the measures. Sixteen tables illustrate the data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306254

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Rater Stringency Error in Performance Rating:

A Contrast of Three Models

Gerald J Cason and Carolyn L Cason

University of Arkansas for Medical Sciences

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Gerald J. Cason
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Abstract

Observed performance ratings are often as much a measure of rater stringency as subject ability. Although activities can be designed to raise the consistency with which raters apply performance rating criteria, practical constraints frequently preclude their use or reduce their effectiveness. Described are 3 models of rating-based performance measurement: Conventional, Handicap, and deterministic Rater Response Theory (RRT). The first offers passive control, the others active control of stringency error. General model requirements, power, bias of measures, computing cost and complexity are contrasted. Some contrasts are illustrated by application of the models to a small and a large set of clinical ("ward") performance ratings of Junior Year US medical students in an Internal Medicine clerkship: 1 and all rotations in an academic year. Both 1- and 2-factor performance domains are considered. Removing stringency error lowered the correlation between domain scores. RRT is the most powerful of the 3 models contrasted. Handicap is a close second, although biased. Under frequently found conditions, RRT obtains any specified inter-rater reliability with only one-third the independent ratings needed by the Conventional model (i.e., mean of observed). Pass/Fail decision outcomes for each model's measure of performance are illustrated for the 163 students in the academic year analyzed and a hypothetical student: Will E. Makit. Willie's fate reflects his true ability, the measurement model, passing score, and stringency and number of his raters. Choice of measurement procedures should reflect a reasoned and reasonable balance in costs of improving the accuracy of the measures and costs of faulty decisions based on those measures.

Correspondence to:

GJ Cason, PhD
UAMS-OED-595
4301 West Markham
Little Rock, AR 72205

Phone: (501) 686-5720

TM013048

Rater Stringency Error in Performance Rating:

A Contrast of Three Models

Gerald J Cason and Carolyn L Cason

University of Arkansas for Medical Sciences
Little Rock, Arkansas 72205

Quantitative ratings based on the judged quality of performance of students, residents, interns, certification candidates, and practicing professionals observed in actual or simulated practice settings provide one of the most common methods for the measurement of ability and competence in the professions. More often than not, there are practically important differences in the stringency with which different raters apply the performance criteria (Cason, Cason & Redland, 1988; Cason, Cason & Stritter, 1986; Cason & Cason, 1984; Delk, Cason, & Reese, 1985; Littlefield, Ellis, Cohen, & Herbert, 1984; O'Donohue & Wergin, 1978). If who rates whom varies from subject to subject (e.g., student to student), differences in rater stringency can cause significant errors in the measurement of subject ability. There are two things which reduce differences in rater stringency: recent participation in the formal development of a rating inventory (e.g., Stiggins,1987) which the developers then use as raters (e.g., the behaviorally anchored scales in Cason, Cason, Bond, & Jackson, 1989) or intensive training in the consistent (within and across raters) use of an extant rating inventory (Stillman,1980). Cost, scheduling conflicts, and other practical constraints usually preclude effective use of either of these approaches to standardizing raters' stringency before ratings are collected. The same constraints also tend to preclude having all subjects rated by the same rater(s). However, since stringency errors are systematic (rather than random), they are amenable to post-hoc mathematical remedies.

This paper is addressed to the use of three remedies for errors in the measurement of ability which arise from differences in rater stringency. The emphasis is upon the contrast between conventional practice and two after-the-fact remedies; that is, models which provide stringency off-setting adjustments to ratings after the ratings have been collected. Since differences in stringency give rise to measurement errors only when raters vary from subject to subject, this very common problematic circumstance is assumed in the following discussion. Although the illustrations in this paper are drawn from medical education, the remedies themselves are broadly applicable. The intent of this paper is pragmatic rather than theoretical: to assist those who use performance ratings in educational, licensing, or certification decisions to make higher quality (i.e., more valid and reliable) decisions. Thus, the examples in this paper are intended only to illustrate and clarify, not prove, the points presented in the discussion.

Models for Interpreting Clinical Ratings

In the following discussion, for convenience the person doing the rating is called the rater and the person whose performance is rated is called the subject. It is assumed that for each subject evaluated by a rater a numeric rating is assigned to each of the one or more criteria on a written inventory of performance

criteria. In spite of evidence to the contrary (C. Cason, Cason & Littlefield, 1983; Dielman, Hull, & Davis, 1980; Keck, Arnold, Willoughby, & Calkins, 1979; Maxim & Dielman, 1987;), to simplify the initial discussion, a single factor is assumed to underlie performance ratings. Thus, an average across ratings on individual criteria is a good measure of a rater's evaluation of a subject's over-all performance. This single value for a rater-subject pair is the rating to which the following discussion refers. (Later in the discussion the issue of multiple factors will be considered.) The basic psychometric assumptions are made: (a) the observed rating (x) is composed of both a true score (t) component and a random error (e_1) component; and, (b) random errors are normally distributed and uncorrelated with true scores. In addition, there may be (and frequently is) a systematic error component (e_2) arising from differences in rater stringency (Wherry, 1952). Thus,

$$x = t + e_1 + e_2 \qquad (1).$$

Conventional Practice Model

In conventional practice, a subject's score is defined as the average (mean) of ratings received from those (one or more) who rated this subject. While the model implicit in this practice makes no special provision for systematic rater error (e_2), the cumulative effect of this error is diminished to the extent that the systematic errors of the raters of one subject become more nearly equal to those for other subjects. That is, rater error is equal to an additive constant for each rater. To the extent that the sum (or average) of the constants for the raters of one subject approaches equality with the sums (or averages) for other subjects, the differential effect of e_2 approaches zero. In other words, as the raters of each subject become more representative (in terms of systematic rater errors) of all the raters in the pool, the net differential effect from subject to subject of such systematic errors tends to "balance out" and approach zero. This kind of representativeness of raters is more likely to be achieved by random assignment of raters to subjects and using large numbers of raters per subject. Random assignment and large numbers of raters per subject tend to reduce the net contribution of both random and systematic error to a subject's observed mean rating. However, as is implied by the formula for the standard error of a mean (see Appendix A, Formula 1), the sampling error of the mean rater error is likely to be large for an individual subject unless the number of raters per subject is large. This means the errors are unlikely to balance out very well when the number of independent raters per subject is small. Also, it must be remembered that even where the number of raters is large, random assignment only tends to produce these balancing effects over-all, not guarantee them in each subject's case. Nevertheless, the over-all balancing out of errors as the number of raters per subject increases is clearly shown by the increasing reliability of the mean of these ratings as estimated by the Spearman-Brown expansion formula (see Appendix, Formula 2).

Spearman-Brown's formula is applicable whether random assignment is used or not. However, if random assignment is not used, systematic rater errors will not tend to be balanced. This will be reflected in a lower estimate of the reliability of a single independent rater and consequent lower total gain in reliability from the multiple independent ratings.

In so far as a justification is possible, this relationship expressed by the Spearman-Brown formula justifies the conventional and wide-spread practice of

ignoring systematic differences in raters' stringency on the grounds that in the long run such differences balance out. If the measure involved is something like a cumulative grade point average, in a program with a fixed curriculum (i.e., no electives, or nearly none), as is the case in many educational programs for the professions, the logic appears reasonably sound for decisions such as choosing the valedictorian. However, this passive approach leaves what to many is an unnecessarily large role to luck or chance. Since the difference in the top few students' GPAs is likely to be well within the standard error of the GPA, one could reasonably wonder if the final ranking of these top students did not as much reflect slightly different luck in the draw of raters as in true ability. Furthermore, in the professions performance in a single course, or very small set of courses, or a single practical exam, may be the basis for major decisions affecting an individual's career: licensure, admission to advanced or specialty training, certification, etc. Here the run is not long enough to have confidence that a passive approach will provide a balanced measure of the subject's ability. It is this question which motivates the use of active remedies for differences in rater stringency provided by the next two models.

Rater Handicap Model

The rater handicap model (Delk, et al., 1985; Littlefield, et al., 1984; Wherry, 1952) is based on the attractively simple notion that if subjects are randomly assigned to raters, and raters each rate a sufficiently large number of subjects, then differences in the mean ratings given by each rater reflect primarily differences in rater stringency rather than differences in the subjects' ability. That is, if assignment is random, and the group assigned to each rater is large, then the mean true ability of subjects in each group should be very little different from the mean in any other group or from the grand mean (i.e., the mean of group means). The formula for the standard error of the mean (Formula 1 in Appendix A), with N the number of subjects per rater, suggests the size of variations in means that would reasonably be attributed to sampling fluctuations. A pair of means more than 2 standard errors apart is unlikely to be solely due to chance ($p<.05$). The handicap model attempts to off-set or correct for systematic differences in rater stringency by computation of a handicap for each rater which is then applied to each rating given by that rater. A rater's handicap is defined:

$$H_i = M_g - M_i, \qquad (2)$$

where:
H_i = rater i's handicap,
M_g = grand mean of all raters' means, and
M_i = mean of rater i's ratings.

The entire difference between the rater's mean and the grand mean is assigned to the rater's stringency even though some part of this may actually reflect random sampling error in the mean true ability of a rater's group of subjects. How good an estimate of the rater's true stringency this may be is a function, in part, of the sampling error of the mean. As mentioned above, the formula for computing a standard error of the mean indicates that small numbers of subjects per rater (e.g., $N_{s/r}<5$) are unlikely to give useful estimates of rater stringency. Furthermore, in only one special case will the handicaps be unbiased least-squares estimators of the raters' stringencies: when the group of subjects rated by each rater is entirely different from that rated by any other rater. That is, no subject is rated by more than one rater. (This is an especially important case

because it is one in which the next model cannot be applied.) In all other cases, the handicaps are only biased approximations of least-squares estimates of stringencies, with the approximation in general improving with increased $N_{s/r}$. (The handicap model may be understood as an approximation of a regression model containing "dummy vectors" for each rater and subject. The handicap-adjusted mean is an approximation of the regression predicted (y') mean for a subject rated by all the raters. The regression equation permits this prediction even though no subject was in fact rated by all raters.)

A subject's handicap-adjusted score is computed by adding each of his or her raters' handicaps to their respective ratings, then finding the average of these. Equivalently, a subject's mean observed rating may be adjusted by adding the mean of his or her raters' handicaps. How much better a measure of the subject's ability the handicap-adjusted score is than the mean of the observed ratings depends directly on how good the handicaps are as estimates of the raters' stringencies.

The great strength of the handicap model is its conceptual and computational simplicity. It can be applied with only modest effort to small data sets with no more computing support than a pocket calculator, and to large data sets with only a spread-sheet program and a micro-computer (Delk, et al., 1985; Mills, 1988; Quattlebaum & Sperry, 1988). One of the costs of this simplicity is that the model provides no intrinsic mechanism for estimating (a) how well the handicap model fits the observed data, (b) the proportion of variance in the observed ratings that the model attributes respectively to rater stringency and subject ability and thus (c) no direct way to estimate the reliability of the adjusted ratings.

For the purposes of this paper, over-all fit (R) and the variance components were estimated using regression analysis. The criterion variable was the observed rating, the predictor variables were rater handicaps and subject handicap-adjusted mean scores. This approach produces an over-all R^2 that slightly over-estimates the model's fit to the data. The distortion appears to be slight. Also, as this is the approach used in the next model it permits more directly comparable results. The stringency (rater effect) and ability (subject effect) components of variance are given by the product of each variable's correlation with the criterion and the variable's beta weight in the regression equation (see Appendix A, Formula 3).

The variance due to subject ability is equal to the intra-class, inter-rater correlation or the reliability of the observed rating of one rater (Ebel, 1951; Hays, 1963). That is, it is the ratio of the variance due to ability versus ability plus rater and random error. Because the variance due to error in the observed includes both the over-all residual error from the regression analysis ($1 - R^2$) and variance due to raters, the estimate of inter-rater correlation (single rater reliability) for handicap-adjusted scores must be higher than that for observed ratings if any variance is attributed to the rater stringency effect. Adjusted ratings have had the effect of rater error removed from them, therefore it is not included in the denominator of the formula. This estimate of reliability for the adjusted scores is, at best, an upper-bound limit for the Handicap model. It is probably an over estimate because it is dependent on the accuracy of the estimated variance associated with raters and the handicap model is more likely to over- than under-estimate this variance. In cases where subjects are rated by multiple independent raters, the Spearman-Brown formula is used to estimate the reliability of the observed means across raters and handicap-adjusted means for subjects.

Deterministic One-Parameter Rater Response Theory (RRT)

Our rater response theory (Cason & Cason, 1984; 1985; 1988) was initially developed in response to the need for a method to off-set systematic rater errors in the assessment of students' clinical performance in health professions educational programs. A method was needed that avoided the rater handicapping model's requirements for random assignment of subjects and relatively large numbers of subjects per rater because often in practice one or the other could not be satisfied. At the level of latent-trait meta-theory, our RRT shares some fundamental notions with and is a conceptual (although not formal mathematical) derivative of the discrete-state, probabilistic item response theories (IRTs) of Lord and Novick (1968) and the so-called Rasch model (Linacre,1989; Rasch,1966; Wright & Stone, 1979). Our RRT is a deterministic rather than probabilistic theory. Our RRT assumes continuous, interval-level rating data and requires the far less complex mathematics of fairly elementary algebra.

Although in the general formulation of our RRT (Cason & Cason, 1984) provision is made for several rater characteristics, here we will be concerned only with our simplified, one-parameter model which we have used in all of our applied work to date. (The expression "one-parameter" follows IRT convention of describing the model in terms of how many parameters are used to represent the measuring device: the item in IRT and the rater in RRT.) The single rater parameter is stringency. In this simplified RRT model illustrated in Figure 1, the observed rating (x) is a curvilinear function of the subject's true ability ($t_{subject}$), the rater's true stringency (t_{rater}) and random error (e). All *systematic* variation in observed ratings is a function of the shape of the Rater Characteristic Curve (RCC) and the difference between the rater's stringency and the subject's ability. The two "s-shaped" (ogival) curves in Figure 1, are Rater Characteristic Curves for raters A and B. The rater's stringency (t_{rater}) is equal to the value of the point on the true ability and stringency (t) scale directly below the point on the RCC associated with a rating half way between the rater's effective rating floor and ceiling. In Figure 1, for rater A, t_{rater} = K; for rater B, t_{rater} = L. In the simplified RRT, this is the point associated with a rating of 50% (of the possible points on the rating inventory); that is, where the rater's stringency equals the subject's ability (t_{rater} = $t_{subject}$) the expected rating is 50%. As implied by Figure 1, the simplified RRT assumes RCCs of all raters have equivalent shapes, i.e., equal slopes and effective floors and ceilings (i.e., 0% and 100%, respectively). Figure 1 also shows that for a subject with ability equal to s ($t_{subject}$ = s) the less stringent rater A is expected to give a higher rating (RA) than the more stringent rater B, who is expected to give the lower rating (RB). More generally, for a fixed ability, the expected rating declines as rater stringency rises. For fixed rater stringency, the expected rating rises as ability rises.

For the RRT to be of practical use, a specific mathematical function must be chosen to stipulatively define the RCC. We arbitrarily chose the cumulative normal ogive to define the deterministic function in our RRT. In the context of our model the normal ogive is simply a function defining an s-shaped curve, it is not a probability distribution function. (Exploratory work with other ogival functions, such as a rescaled, translated, inverse-cotangent function, have not given quite as good fit with the empirical data.)

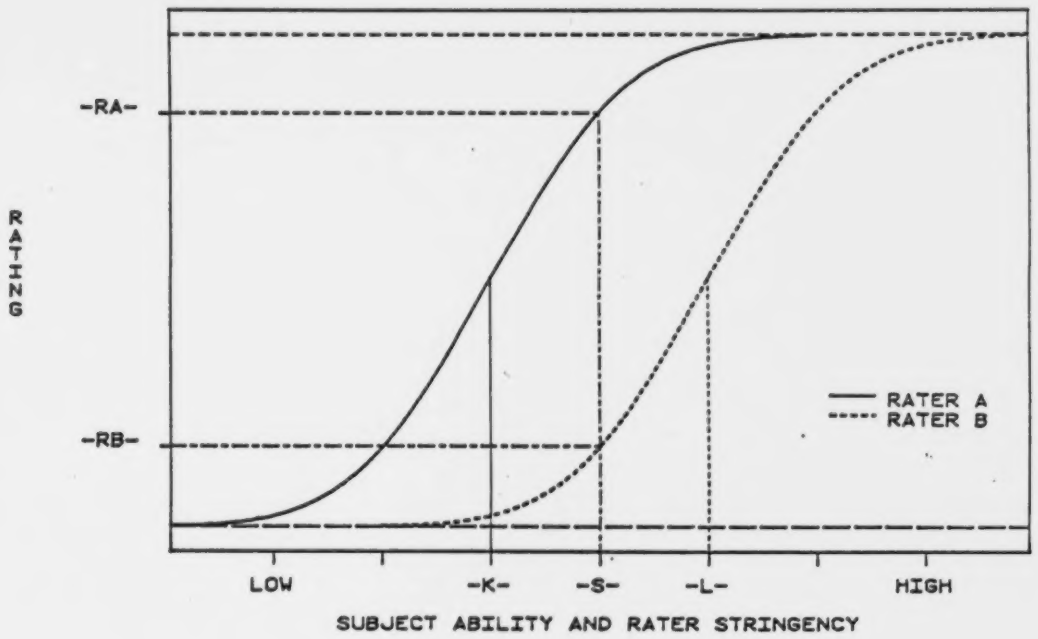

Figure 1. Rater Characteristic Curves for Raters K and L.

Expressing the observed ratings (x) as proportions, the RRT model is formally defined[1]:

$$x = \{ z [(t_{subject} - t_{rater}) / a] \} + e. \qquad (3)$$

where:
z = unit-normal (cumulative function) deviates,
and
a = 100; an arbitrary scaling factor.

Unlike common practice in IRT which results in the use of small and negative values to represent subject abilities and item difficulties, we introduce a large, positive constant scaling factor (a) and choose the t-scale origin (by arbitrarily defining some rater's t_{rater} = 500) to obtain values for ability ($t_{subject}$) and stringency (t_{rater}) that are large positive numbers. Since, Formula 3 expresses a curvilinear relationship, illustrated in Figure 1, it does not lend itself to the more efficient, economical numeric analysis procedures available on computers, e.g., linear regression, for the estimation of the model's parameter values (i.e., true abilities and stringencies) from actual observed data. However, the inverse unit-normal (z^{-1}) function may be applied to the observed ratings if they are expressed as proportions:

---

[1] In earlier work the symbol RRP was used for t_{rater} and SAP for $t_{subject}$.

$$y = z^{-1}(x). \qquad (4)$$

By the transformation of x to y, Formula 3 may be converted into a form amenable to linear regression analysis:

$$y = [\,(t_{subject} - t_{rater}) / a\,] + e. \qquad (5)$$

The details of parameter estimation are given elsewhere (Cason, Cason & Redland, 1988a, 1988b; Cason & Cason, 1985) and are beyond the scope of this discussion. Suffice it to say that the process involves the application of linear regression analysis to a criterion (y) variable defined by z^{-1}-transformed observed ratings and predictors defined by binary coded "dummy" variables for each rater and subject. The model parameters are found by a linear transformation of the raw regression weights associated with raters and subjects. As discussed above, results of the regression analysis and the appropriate formula (in Appendix A) provide estimates of over-all fit (R), components of variance, reliabilities, and standard errors. (The details of the application of these formula to RRT analysis are given in CL Cason, et al., 1988a.)

No use of the means or other distributional characteristics of observed scores, stringencies, or abilities was made in either the definitions or derivations of the defining Formula (3, 5) of the RRT. Unlike the Rater Handicap model, random assignment of subjects to raters is not required to obtain good estimates of the RRT's parameters. (The only distributional assumption in the RRT, is that random errors are normally distributed on the t-scale. This implies, given the curvilinear relationship of the t-scale to observed ratings, that errors are not normally distributed on the observed rating scale.) Although, random assignment is not required by the RRT, it has another special requirement.

Disregarding scaling constants, RRT basically asserts observed ratings reflect distances between raters and subjects on the t-scale. If a solution for abilities and stringencies is to be obtained without any use of rater or subject means, the data must be "coupled" in a particular manner. The data must provide information sufficient to find the distance from any rater to any subject. That is, there must be a path that leads from any subject to any other subject or rater; and conversely from any rater to any other rater or subject. This condition is not as hard to satisfy as it may at first appear. All the subjects rated by rater A are coupled to A, that is observed data is available about their distances from A. If one of these, subject J, is also rated by rater B, then rater B and all the subjects rated by rater B are coupled to rater A through subject J. If there are over-laps in the subjects of different raters, this additional information is used to find unbiased estimates of the true distances. Regression analysis is one procedure which does this.

It should be intuitively apparent that even if the subject by rater table (matrix) has many empty cells (i.e., each rater rates only a few subjects and each subject is rated by only a few raters), the coupling requirement can nevertheless easily be more than minimally satisfied. As the number of alternate (overlapping, redundant) pathways for which observed data provides distance information increases, the accuracy of stringency and ability estimates improves. In addition to satisfying the coupling requirement, as a minimum the data should provide at least two ratings per subject and two ratings per rater. Figure 2 provides a schematic illustration of a rating data matrix that satisfies both of these RRT requirements.

	Rater				
	1	2	3	4	5
Subject					
1	x	x			
2		x	x		
3			x	x	
4				x	x
5	x				x

Figure 2. Example of Coupled Data.

RRT-adjusted Rating

In so far as the analysis of the observed ratings using the RRT is successful, the estimates of subject ability expressed on the t-scale ($t_{subject}$) are free from the effects of rater stringency. It could be used as the measure of the subject performance. However, unlike the application of IRT to examination data, in performance rating there is a presumptively appropriate scale printed on the original performance rating inventory. For many purposes, it is best to express the results of the RRT analysis on this original scale rather than on the t-scale which is likely to be unfamiliar to and therefore difficult for the users of the results (e.g., faculty, licensing authorities, etc.). The exact way in which this is done has major evaluative (standard setting) implications. Once the true abilities of subjects and true stringencies of raters are estimated, the expected rating (x') of a subject rated by any of the raters may be calculated as:

$$x' = z\,[\,(t_{subject} - t_{rater}) / a\,]\,. \qquad (6)$$

What rater or group of raters, and if a group whether unequal weights should be applied to the separate estimates in computation of the final RRT-adjusted score is a policy issue. For example, raters with more experience might be given greater weight, or the stringency of the head of a department alone might be adopted as the consistent standard of stringency. In the following examples, the average of a subject's expected ratings from all raters, i.e., the estimate of the mean rating the subject would have received if rated by all the raters, is used as the RRT-adjusted rating.

RRT Rater Handicap

In a similar manner, a rater's stringency is represented by a value on the t-scale. However, the t-scale is not linearly related to the observed rating scale; therefore, the t-scale values of stringencies are not directly comparable to the handicaps for raters used in the Handicap model. For purposes of comparisons between the two models an analog of a rater handicap is computed from the RRT results as follows. Using a rater's t-scale stringency, and each subject's t-scale ability, the rater's expected mean had he rated all subjects is found. These RRT-expected rater means are then used in place of observed means to obtain RRT rater handicaps as described in Formula 2.

Characteristics of Data Required by Each Model

Table 1 is intended to aid practical decisions about which model may be used under various constraints. It provides a summary of the data characteristics required by each of the models (Conventional, Handicap, and RRT), if it is to be used for the control of systematic rater stringency error. All the models do a better job with more data. The minimum number of subjects per rater for the Handicap model given in the table represents our best judgement, based on fragmentary, suggestive evidence. Of the three models, in general, RRT makes the most efficient use of information in performance ratings; the Conventional model makes the least efficient use of the information. If the RRT analysis is applied to data that satisfy the Handicap requirements and only satisfy the coupling requirement within subsets of the data, the results have a mixture of the properties of both models.

Table 1
Comparison of Model Requirements and Characteristics

	Models		
	Conventional	Handicap	RRT
A. To Off-set any Rater Error			
Random assignment	Yes	Yes	No
Data "coupling"	No	No	Yes
Min raters/subject	2	1	2
Min subjects/rater	1	5	2
B. Efficiency/Power	Low	High	Highest
C. Quality of Measures			
Unbiased	Yes	No[2]	Yes
Least-squares	Yes	No[2]	Yes
D. Computing			
Cost	Trivial	Very Low	Low
Complexity	Means	Deviates	Regression
E. Overall Quality	Lowest	2nd best	Highest

Method

The following illustrative examples are based primarily upon the application of the three models to two sets of data, both obtained from University of Texas Health Science Center-San Antonio, Texas. Both data sets meet both the random

---

[2]The handicap model produces least squares, unbiased measures of performance only in the special case where each subject is rated by one and only one rater.

assignment requirement of the handicap model and the coupling requirement of the RRT model. The smaller set is from a single cohort of students; the larger is drawn from the entire academic year. The data all come from the College of Medicine's Junior year clerkship (clinically oriented course) in Medicine for the year 1982. In addition, the data for one student, somewhat whimsically denoted "Will E. Makit" or simply "Willie", who went through the same course more recently (1989), was interpolated into the results to provide the concrete, specific examples. These examples were included to provide a common point of reference between the simulated "grading committee meeting" exercise and the presentations of the contributors to the AERA symposium (Session number 23.16) in which this paper was first presented. The data analysis results reported for the single cohort and whole academic year did not include Willie's data. His data were used only in the illustrations specifically referring to him.

Rating Inventory

Students were rated using an inventory containing 5 criteria. Each criterion was on a 0 to 14 scale. The first 4 criteria were determined by the authors, on the basis of connotative content, to be measures of cognitive and technical skills. The fifth item was the only one in the inventory that appeared to measure affective, inter-personal and communication skills. The data have been used in a number of previous studies (e.g., C. Cason, Cason, & Littlefield, 1983; Cason & Cason, 1985; G. Cason, Cason, & Littlefield, 1983; Littlefield, Harrington, Anthracite, & Garman, 1985).

Dependent Measure

Two kinds of assumptions were made leading to three different definitions for the dependent measure. Assuming that the ratings of clinical performance represented a single general clinical competence factor, a student's observed rating from a rater was defined as the mean of the ratings assigned by this rater to the 5 items on the inventory. Assuming that there were two factors represented by the criteria, the average of the first 4 defined the observed rating for Cognitive-Technical (CT) skill and the score on the fifth defined the observed rating on Affective-Interpersonal (AI) skill.

Raters

The students' clinical performance was rated by Faculty attending physicians and senior residents.

Data

Set 1 consisted of the ratings of students in the first cohort to rotate through the Medicine Clerkship in 1982 (Cohort 1982a). Total number of ratings (inventories completed) was 129. Number of raters was 42. Number of students was 24. Thus, there were 3.07 ratings per rater; 5.38 ratings per subject

Set 2 consisted of all ratings for all students during 1982 (Cohorts a through g combined) excluding those ratings that came from raters who rated less than 5 students during the year. This exclusion did not remove any student from the analysis. It did remove 94 raters who together had given 219 ratings (completed 219 inventories). After exclusion of data from raters with fewer than 5 subjects, there were 744 ratings, by 93 raters, on 163 students. Thus, there were 8 ratings

per rater; 4.56 ratings per student. These data met the minimum requirements of all the models, that is, the Conventional, Handicap and RRT models.

Results

Assuming Uni-dimensionality of Clinical Ratings

Application of the Models to a Small Data Set

Table 2 provides descriptive statistics on the performance measure provided by each of the three models across the 24 students in the cohort. For the Conventional model, the measure is each student's observed mean (across raters); for the other models, it is their respective adjusted scores. All models give very nearly the same mean across students. However as reflected in the standard deviations (SDs) and ranges (minimum - maximum), the Handicap model has a somewhat lower variability than the Conventional model and the RRT model has a much larger variability than either of the others. The lower variability of the Handicap model arises from assuming all the difference between the raters' grand-mean and individual rater means is due to differences in raters. To the degree this is not so, too much is subtracted from the observed scores in the computation of handicap-adjusted scores forcing them all closer to the grand mean than is truly justified. The higher variability of the RRT model directly reflects RRT's implicit assumption that a rater tends to systematically under-estimate the ability of very bright students (by that rater's standard of stringency) and over-estimate the ability of very dull students. Thus, generally RRT will give measures of performance having greater across students variability than the observed scores or handicap-adjusted scores. Under random assignment as number of subjects per rater rises, these values are expected to gradually converge.

Table 2
Cohort 1982A
Descriptive Statistics on Performance Measures from the Three Models

	Models		
	Conventional (Observed)	Handicap (Adjusted)	RRT (Adjusted)
Mean	63.99	63.69	64.12
Minimum-Maximum	45.14-77.14	44.77-72.80	42.45-85.36
Standard Deviation	7.17	5.89	13.29

$N_{raters} = 42$; $N_{subjects} = 24$; $N_{ratings} = 129$

Table 3 summarizes the fit of the three models to the 1982a cohort. For the Handicap and RRT models it also shows the estimated proportions of variance due to raters' stringency and subjects' ability. The Conventional model makes no provision for separate contributions from systematic rater effects and subject ability: thus, no estimates of their separate contributions is applicable to this model. Since for these data, the rater stringency effect is statistically significant ($F = 2.36$; $df = 23, 65$; $p < .0001$) in both the Handicap and RRT models, the Conventional model cannot account for the observed data as well as either of the others. The Handicap model attributes considerably less of the variance in the observed ratings to ability than does the RRT model; and, a little more to

stringency than does the RRT model. The components of variance at the top of Table 3 were used to compute the estimates of reliabilities given at the bottom of Table 3.

Table 3
Cohort 1982A
Fit, Components of Variance, and Reliability
for the Three Models

	Models		
	Conventional	Handicap	RRT
Fit: R	.49	.82	.85
Variance			
Stringency	.00	.49	.45
Ability	.24	.17	.27
Total (R^2)	.24	.66	.72
Reliability			
1 Rater			
Observed	.24	.17	.27
Adjusted	[.24]	.33	.50
Mean of 5.38 Raters			
Observed	.63	.52	.67
Adjusted	[.63]	.73	.84

In this context where there are relatively few raters per subject (Nr/s = 5.38) and few subjects per rater (Ns/r = 3.07) neither the Conventional model nor the Handicap model can be expected to perform very well (i.e., give highly reliable results). The Conventional model relies on passive balancing of the rater effect; the Handicap model needs a higher Ns/r to get good estimates of the rater stringency. By default, RRT becomes the "gold standard" for the interpretation of these results because it is least sensitive to small n (and the kind of problem with large sampling errors caused by small n for the other models). (An empirical rather than logical defense of RRT as the "gold standard" would require something like showing its adjusted scores had greater predictive validity than those from either of the other models.) Therefore, while the estimate of reliability using all the data for each model (i.e., Nr/s) reported in Table 3 indicates that reliability improves from Conventional to Handicap to RRT; Handicap may not be an improvement over Conventional. As is expected from the moderate correlation ($r = .55$; $p \leq .001$) between the rater handicaps used in the Handicap model and their analogs from the RRT model, Table 4 shows only a moderate correlation between the adjusted scores of subjects for these two models. The Handicap model's adjusted scores are more strongly correlated with the least powerful (Conventional) model's scores, rather than with the most powerful (RRT) model's. This is consistent with the Handicap model's expected relative inability to properly apportion the variance to stringency and ability when Ns/r is small. As

Ns/r increases, the correlation between the adjusted scores of the Handicap and RRT models should rise. In the case of very low numbers of students per rater, RRT is best, the Conventional model a distant second, and the Handicap model is not useful.

Table 4
Cohort 1982A
Correlations among Performance Measures from the Three Models

	Models		
	Conventional (Observed)	Handicap (Adjusted)	RRT (Adjusted)
Handicap	.86	1.00	.60
RRT	.65	.60	1.00

N = 24; for all rs $p < .001$.

Application of the Models to a Large Data Set

The results reported in Tables 5 through 7 are from application of the three models to all the 1982 academic year data that satisfy minimum requirements for both the Handicap and RRT models. Compared to the results obtained with the small data set, Table 5 indicates the means are again quite similar and, as expected, the variabilities have converged to near the same value with those from RRT being highest and those from Handicap lowest. Also, RRT and Handicap fit the observed data about equally well over-all (R and R^2 in Table 6), with RRT slightly better. A significant rater stringency effect was again found by both models ($F = 2.84$: $df = 164, 488$; $p < .0001$). Handicap does a better job apportioning the variance between rater stringency and subject ability, but again somewhat over-estimates the rater component and underestimates the subject component. This leads to the pattern of single and multiple-rater reliabilities shown at the bottom of Table 6: RRT-adjusted best, Handicap-adjusted a close second and Conventional a not too distant third. The larger number of subjects helps both the Conventional and Handicap models do a better job, thus their values are converging on those of the RRT model. (In contrasting the reliability results of the large and small data sets, keep in mind the average number of raters per subject is higher in the smaller set, thus the single-rater reliabilities are more directly comparable between data sets than are the multiple-rater reliabilities.) A much higher correlation between rater handicaps used by the Handicap model and their analog from RRT (.85 here, rather than .55 in the smaller data set), is another indicator that Handicap and RRT models are converging. This pattern continues in Table 7: the measures of performance given by the models are more highly inter-correlated in the larger set than in the smaller. That is, the models' measures of performance are converging. Nevertheless, RRT remains the gold standard. Since the Handicap model's measures of performance are now more highly correlated with the gold standard (RRT) than with the Conventional model (which continues to have the lowest correlation with RRT), this illustrates that improved measures can be had through the Handicap model, if the requirements shown in Table 1 are met.

Table 5
Academic Year 1982
Descriptive Statistics on Performance Measures from the Three Models

	Models		
	Conventional (Observed)	Handicap (Adjusted)	RRT (Adjusted)
Mean	65.71	65.71	66.28
Minimum-Maximum	33.57-92.86	37.24-93.04	41.14-98.57
Standard Deviation	9.46	7.80	9.96

N_{raters} = 93; $N_{subjects}$ = 163; $N_{ratings}$ = 744

Table 6
Academic Year 1982
Fit, Components of Variance, and Reliability
for the Three Models

	Models		
	Conventional	Handicap	RRT
Fit: R	.65	.80	.80
Variance			
Stringency	.00	.32	.26
Ability	.43	.32	.38
Total (R^2)	.43	.64	.65
Reliability			
1 Rater			
Observed	.43	.32	.38
Adjusted	[.43]	.47	.52
Mean of 4.56 Raters			
Observed	.77	.68	.74
Adjusted	[.77]	.80	.83

Table 7
Academic Year 1982
Correlations Among Performance Measures from the Three Models

	Models		
	Conventional (Observed)	Handicap (Adjusted)	RRT (Adjusted)
Handicap	.89	1.00	.93
RRT	.87	.93	1.00

N = 163; for all rs $p < .001$.

Model Choice and Pass/Fail Frequencies

It might be tempting to conclude that since the models produce moderately highly ($.87 < r < .93$) correlated measures with about the same means, one is as a good as another. But the three models produce different outcomes for specific individuals and the best decisions will be obtained with the RRT model. Table 8 illustrates the impact of model choice on pass/fail outcomes for the Medicine clerkship students in the 1982 academic year if the minimum passing score had been defined as a rating of 50% (i.e., labeled on the rating inventory: "adequate performance without significant deficits"). (This passing cut-off was chosen simply for illustrative purposes. We have no information on whether it is higher, the same, or lower than that actually used. Minor changes in the cut point can, depending on the distribution of scores, produce dramatic differences in the results.) Not only are the total number who fail different depending on the model used: Conventional (8), Handicap (4), RRT (7), but exactly who fails and who passes varies depending on the model. Since the RRT model provides the best measure of performance, the lower right hand cell of Table 8 shows that the second best measure (Handicap-adjusted) would have led to probably unjustified passes for three students. The Conventional model would pass 3 who probably should have been failed, and fail 4 who probably should have passed (upper right hand cell of Table 8).

Table 8
Academic Year 1982
Transitions in Pass/Fail

	Decision Based on			
Decision Based on	Handicap Model		RRT Model	
Conventional Model				
	Pass	Fail	Pass	Fail
Pass	160	1	159	3
Fail	5	3	4	4
Handicap Model				
Pass			159	3
Fail			0	4

Willie's Fate

As shown in Table 9, Student Will E Makit's fate, had he been in this class, would have depended upon the model, the passing score adopted, and what the "luck of the draw" gave him as raters. Had Willie's observed ratings come from a group of easy raters (a group whose mean RRT true stringency was 1 standard deviation below the mean across all raters), the Handicap and RRT models would fail him, although RRT by just one point. The numbers recorded on Willie's rating sheet by easy raters over-state his competence relative to what other raters would have assigned for the performance. Had Willie's moderately low ratings been given by a group of stringent raters (mean RRT true stringency +1 SD above the over-all rater mean), the observed ratings would have understated his true ability. If his raters had been of middling stringency, he passes under all models. If the passing score had been set as high as 60%, and his low observed ratings had come from stringent raters, his only chance for fair treatment would have been under one of the models that actively adjusts for rater stringency.

Table 9
Willie's Fate in the Course
P = Pass; F = Fail

If raters had been	Then Willie would		
	Conventional	Handicap	RRT
Easy	P (50.4)	F (43.3)	F (49.0)
Middling	P (50.4)	P (50.4)	P (51.9)
Hard	P (50.4)	P (60.5)	P (60.3)

Assuming Two Factors Underlie Clinical Ratings

In general, the pattern of results across models within each performance domain (i.e., Cognitive and Affective) are parallel to each other, and to the results based on assuming a single underlying global factor (presented above) with respect to fit, components of variance, and reliabilities. Therefore, the results for the two-factor case will not be presented or discussed in the same detail. Table 10 shows that within domain, all models give about the same mean for Cognitive (69 to 70) and for Affective (65) performance. Thus, regardless of model these students were judged to be slightly better in their Cognitive performance. As with the global measure (i.e., single factor assumption), RRT produced the largest and Handicap the smallest variability in each domain. Table 11 reveals the same general pattern of fit, distribution of variance, and consequent relationships among reliability estimates as was seen in the global measure. Within domain, RRT fits best and gives the highest estimated reliability. However, all models fit the Affective domain slightly better than they did the Cognitive domain. This is contrary to what might be expected since the measure of Cognitive performance was based on the mean of 4 inventory items, while for Affective performance only a single item was used. The mean of 4 items might be expected to contain less random error than the rating of a single item. This expectation is not supported.

Table 10
Academic Year 1982
Descriptive Statistics for Two Domains from the Three Models

	Models		
	Conventional (Observed)	Handicap (Adjusted)	RRT (Adjusted)
Cognitive-Technical Performance			
Mean	68.75	69.08	70.46
Minimum-Maximum	35.72-90.48	41.51-92.36	44.60-98.42
Standard Deviation	9.02	7.37	11.00
Affective-Interpersonal Performance			
Mean	64.95	64.87	65.43
Minimum-Maximum	32.15-93.45	36.18-93.21	38.82-98.62
Standard Deviation	9.87	8.21	10.39

Table 11
Fit, Components of Variance, and Reliability
Cognitive-Technical and Affective-Interpersonal Performance

	Cognitive-Technical			Affective-Interpersonal		
	Conv	Hcap	RRT	Conv	Hcap	RRT
Fit: R	.55	.77	.77	.66	.80	.80
Variance						
Stringency	.00	.35	.34	.00	.30	.25
Ability	.30	.25	.25	.44	.34	.40
Total (R^2)	.30	.60	.59	.44	.64	.65
Reliability						
1 Rater						
Observed	.30	.25	.25	.44	.34	.40
Adjusted	[.30]	.38	.38	[.44]	.49	.53
Mean of 4.56 Raters						
Observed	.66	.61	.61	.78	.70	.75
Adjusted	[.66]	.74	.74	[.78]	.81	.84

Also consistent with the pattern for the global measure, Table 12 shows a pattern of correlations of performance measures in which the adjusted measures (Handicap and RRT) are more highly correlated with each other than with the less powerful Conventional model for the Affective domain. However, for reasons not immediately apparent, the pattern was not sustained in the Cognitive domain.

Table 12
Correlations Among Performance Measures from the Three Models

	Models		
	Conventional (Observed)	Handicap (Adjusted)	RRT (Adjusted)
Cognitive-Technical Performance			
Handicap	.89	1.00	.87
RRT	.80	.87	1.00
Affective-Interpersonal Performance			
Handicap	.90	1.00	.93
RRT	.88	.93	1.00

N = 163; for all rs $p < .001$.

Model Choice and Pass/Fail Frequencies

The differences in the mean performance scores for each domain (shown in Table 10) imply that passing frequencies might not be equal, regardless of model, in both domains. However, differences in the means provide a chancy guide to what will occur in the extreme tails of the distribution of scores. Continuing to assume that a minimum passing score is 50%, Table 13 gives the over-all pass and fail frequencies that would have been observed in the two domains during the 1982 academic year. As discussed above, the particular students that pass or fail depend on the model, the pass criterion, the domain, and, in addition, how the information from the two domains is combined. Low reliabilities and high intercorrelations among sub-scale scores in most settings render them useless as independent bases for evaluation. While not as high as might be desired for individual assessment (reliability >.95), the reliability for the RRT measures of each domain are probably minimumly acceptable for assigning "grades" in a single clinical course. However, are they measuring sufficiently different aspects of performance to be treated as independent summary assessment categories? The pattern of correlations in Table 14, suggests that they are not. Under the Conventional model, the stipulative measures of the two domains have a very nearly perfect correlation. Although this drops to the relatively moderate value of .63 when rater stringency error is removed by the RRT model, this is still too high to unequivocally show that separate factors have indeed been measured. This suggests two things. As been the case in other studies (Forsythe, McGaghie, & Friedman, 1986), the method for defining the factors here was not sufficiently powerful. However, if RRT were combined with factor analysis, the removal of the

stringency error might very well allow the factor analysis to find a stronger factor structure than has been thus far possible.

Table 13
Pass/Fail Frequencies by Domain and Model

	Model					
	Conventional		Handicap		RRT	
	CT	AI	CT	AI	CT	AI
Pass	156	155	159	159	160	155
Fail	7	8	4	4	3	8

CT = Cognitive-Technical Performance
AI = Affective-Interpersonal Performance

Table 14
Intercorrelations among Measures in the Two Performance Domains from the Conventional and RRT Models

	Model			
	Conventional		RRT	
	CT	AI	CT	AI
Conventional:				
CT	1.00	.99	.63	.87
AI		1.00	.62	.88
RRT:				
CT			1.00	.63
AI				1.00

CT = Cognitive-Technical Performance
AI = Affective-Interpersonal Performance

Willie's Fate

Even though it is doubtful that good measures of the two domains were obtained, for illustrative purposes Student Will E. Makit's case will be treated as if we had. Willie's fate on both factors is depicted in Table 15. The raters are the same as those used in the global example. That is, they are lenient (easy graders), middling, or stringent (hard graders) in over-all terms. A particular individual rater might be equally or differentially demanding in the two domains. If raters were chosen to be easy or hard within each domain, then the example would simply mirror the results for Willie on the global measure within each specific domain (with some differences in the exact values of the ratings obtained).

Table 15
Willie's Fate
Pass = 50% in each Domain

If raters had been	Then Willie would					
	Conventional		Handicap		RRT	
	CT	AI	CT	AI	CT	AI
Easy	F(49)	P(63)	F(42)	P(54)	F(39)	F(49)
Mid	F(49)	P(63)	F(49)	P(63)	F(48)	P(64)
Hard	F(49)	P(63)	P(59)	P(74)	P(58)	P(75)

The thing to note about Willie's fate is the impact of the RRT model versus the Conventional model, and the over-all stringency of the raters who gave him the observed scores. The Conventional model assumes all the students had equally stringent raters (or that the differences balanced out) so Willie fails the Cognitive domain and passes the Affective regardless of the true stringency of his raters. In the RRT measures, the outcome is the same as the Conventional model only in the "middling" case; i.e., where the Conventional model's assumption is correct. If his raters were easy, his observed ratings over-stated Willie's accomplishments in both domains. In fact in this example, they were most over-stated in the Affective domain. Willie's RRT-adjusted Affective score was 14 points below his observed score; his Cognitive score falls only 10 points. As a result he fails on both domains. Had Willie's observed ratings come from a group that over-all was hard, using RRT-adjusted scores he would have clearly passed both domains. The Handicap model produces out-comes between the others: like Conventional for easy raters, like RRT for hard raters.

If separate passing scores are required to pass the clerkship, some interesting questions are raised. Is education relevant to each domain given? Most clerkships are oriented more toward teaching the Cognitive domain. In this case, it makes sense to have Willie repeat the clerkship if his failure was Cognitive. But what if it were both domains, or only Affective? Is it more a matter that the course gives the student an opportunity to discover what works, to teach himself as it were, rather than be taught? If so, it might reasonably be required that Willie repeat the experience until he figures out how to perform up to standards in both domains, gives up or runs out of opportunities.

Conclusions and Recommendations

Even in those cases where effective consensus building procedures can be used with raters, there is likely to remain some systematic rater error. Each of the three models discussed above can, if properly employed, reduce this residual error. However, each model has its own requirements and costs. In general, there is a need to collect and analyze performance rating data in a manner that shows the extent to which systematic rater error is present. Without such information, it is no more than a guess what is needed in a specific case. Absent this specific information, Table 16 provides general guidance based on circumstances that are fairly typical where consensus building procedures are absent or unsuccessful. The values in Table 16 for the number of raters per student for each model to achieve a certain reliability assume that the model's requirements (Table 1) are met and the components of variance for rater stringency and subject ability are .30 and .40, respectively. Given these

assumptions, the Conventional model requires about 3 times as many independent raters per student to achieve the same level of measurement accuracy. The rightmost column in Table 16 shows the percentage of reversals in rank that would be expected for a pair of individuals initially at the 75th and 50th percentile, upon re-examination. This does not mean their percentile rankings would reverse, but simply that on second evaluation, the student you thought was in the top quarter would obtain a score below the student you thought was in the 3rd quarter of the group evaluated. How troublesome errors of this magnitude are will suggest the level of reliability you need. How much it costs will probably decide what you have to accept.

Table 16
Percent of Reversals of Rank on Reassessment
for Students at the 75th and 50th Percentile

Number of Raters/Student Needed				Percent
Conventional	Handicap	RRT	Reliability	Reversals
5.44	1.75+	1.75	.70	27.10
9.33	3.02+	3.02	.80	19.70
21.00	6.79+	6.79	.90	8.70
44.33	14.33+	14.33	.95	2.20
114.33	36.96+	36.96	.98	0.05

Assumes rater variance = .30; subject variance = .40.
Adapted, in part, from Thorndike and Hagen (1977).

Cost of additional raters, or rater training are likely to be far greater than those of using the Handicap or RRT models. Since the RRT model is most powerful, it would seem reasonable to begin by applying it. If the data structure does not meet RRT's requirements, use the Handicap model. If the structure doesn't meet either; then change it, such that it does. If that can't be done, its more raters or rater training. And if raters can't be assigned randomly, rater training or some other consensus building procedure is the only avenue left.

References

Cason, C.L., Cason, G.J., & Littlefield, J.H. (1983). Variation in intra-rater stringency in cognitive-technical and affective-interpersonal clinical performance domains. Presented at the annual meeting of the American Educational Research Association, April.

Cason, C.L., Cason, G.J., & Redland, A. (1988). Off-setting differences in reviewer stringency. Abstract in *Resources in Education*, *23*(7), 142. ERIC Document ED 287-888.

Cason, C.L., Cason, G.J., & Redland, A. (1988). Peer review of research abstracts. *Image: Journal of Nursing Scholarship*, *20*(2), 102-105.

Cason, C.L., Cason, G.J., & Stritter, F.T. (1986). Reviewer stringency and proposal quality in the selection of the 1986 AERA Division I program. *Professions Education Research Notes*, *8*(1), 7-8.

Cason, C.L., Cason, G.J., Bond, M.L., & Jackson, E. (1989). Defining, measuring, and evaluating CNS clinical performance: Technical assessment of reliability and validity. Paper presented at the 7th Annual Conference on Research in Nursing Education, National League for Nursing, January.

Cason, G.J., & Cason, C.L. (1984). A deterministic theory of clinical performance rating: Promising early results. *Evaluation & the Health Professions*, *7*, 221-247.

Cason, G.J., & Cason, C.L. (1985). A regression solution to Cason and Cason's model of clinical performance rating: Easier, cheaper, faster. Paper presented at the annual meeting of the American Educational Research Association, April. *Resources in Education*, 1986, *21*,147 (Abstract) ERIC Document ED 262-079

Cason, G. J., & Cason, C.L. (1988). Improving peer review of nursing research. *Computers in Nursing*, *6*(6), 253-262.

Cason, G.J., Cason, C.L., Littlefield, J.H. (1983). Controlling rater stringency error in clinical performance rating: Further validation of a performance rating theory. Abstract in *Resources in Education*, *18*(8), 176. ERIC Document ED 228-314

Delk, J.L., Cason, G.J., & Reese, W.G. (1985). A practical method to enhance fairness of clerkship ratings. *Journal of Medical Education*, *60*, 944-945.

Dielman, T.E., Hull, A.L., & Davis. W.K. (1980). Psychometric properties of clinical performance ratings. *Evaluation & the Health Professions, 3*, 103-117.

Ebel, R.E. (1951). Estimation of reliability of ratings. *Psychometrika*, *16*, 407-424.

Forsythe, G., McGaghie, W., & Friedman, C. (1986). Construct validity of medical clinical competence measures: A multitrait-multimethod matrix study using confirmatory factor analysis. *American Educational Research Journal*, *23*, 315-336.

Hays, W.L. (1963). *Statistics.* New York: Holt, Rinehardt, & Winston.

Keck, J.W., Arnold, L., Willoughby, L., & Calkins, V. (1979). Efficacy of cognitive/noncognitive measures in predicting resident-physician performance. *Journal of Medical Education*, *54*, 759-765.

Linacre, J.M. (1989). Objectivity for judge-intermediated certification examinations. Paper presented at the annual meeting of the American Educational Research Association, AERA, March.

Littlefield, J.H., Ellis, R., Cohen, P. & Herbert, R. (1984). Leniency and score distribution differences among clinical raters. *Research in Medical Education: Proceedings of the Twenty-Third Annual Conference.* Washington, DC: Association of American Medical Colleges.

Littlefield, J.H., Harrington, J.T., Anthracite, N.E., & Garman, R.E. (1985). A description and four-year analysis of a clerkship evaluation system. *Journal of Medical Education, 56*, 334-340.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison Wesley.

Maxim, B., & Dielman, T. (1987). Dimensionality, internal consistency and interrater reliability of clinical performance ratings. *Medical Education*, ,*21*, 130-137.

Mills, M.L. (1988). A method to calculate and analyze resident's evaluations by using a microcomputer data-base management system. *Journal of Medical Education*, *63*, 724-727.

O'Donohue, W.J., & Wergin, J.F. (1978). Evaluation of medical students during a clinical clerkship in internal medicine. *Journal of Medical Education, 53*, 55-58.

Quattlebaum, T.G., & Sperry, J.B. (1988). A computerized system for evaluation of residents and residency experiences. *AJDC, 142*, 758-762.

Rasch, G. (1966). An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology, 19*, 49-57.

Stiggins, R. (1987). NCME instructional module on design and development of performance assessments. *Educational Measurement: Issues and Practice, 59*, 33-42.

Stillman, P.L. (1980). Arizona clinical interview medical rating scale. *Medical Teacher, 2*, 248-251.

Thorndike, R.L., & Hagen, E.P. (1977). *Measurement and evaluation in psychology an education* (4th ed.). New York: Wiley.

Wherry, R.J. (1952). The control of bias in rating: A theory of rating (Personnel *Research Report 922)*. Washington, DC: Department of the Army Personnel Research Section, February.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago: MESA Press.

Appendix A

Formula 1: Standard Error of the Mean

$$SE_{mean} = SD_x / Sqrt(N_{raters\ per\ subject})$$

where:
SD_x = standard deviation of the measure (x).
Sqrt(n) = square root of n

Formula 2: Spearman-Brown Expansion Formula

$$r_{kk} = (k * r_{xx}) / \{ 1 + [(k - 1) * r_{xx}] \}$$

where:
r_{kk} = reliability of mean of k independent ratings,
r_{xx} = reliability of k = 1 independent rating (equivalent to the inter-rater correlation),
k = Nr = number of independent ratings per subject.

Formula 3: Computing the components of variance

$$v_a = r_{ay} * b_a.$$

where:
v_a = proportion of variance attributable to a,
r_{ay} = correlation between variable a and criterion y, and
b_a = standardize regression weight (beta) of variable a.

Formula 4: Intra-class correlation

$$r_{ic} = v_a / (v_a + v_e)$$

where:
r_{ic} = intra-class correlation,
v_a = variance due to effect a, and
v_e = variance due to error,

In the Handicap and RRT models, v_e = rater variance and residual error ($1-R^2$) for observed means; however, v_e = only residual error ($1-R^2$) for the adjusted score.